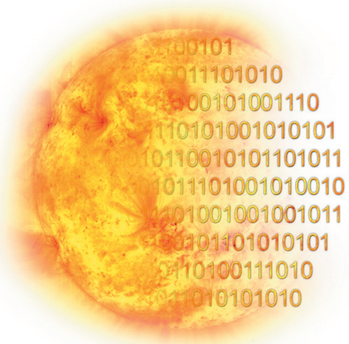
IN31B-1008

Ignored Issues in e-Science: Collaboration, Provenance and the Ethics of Data

Joseph A. Hourclé, NASA GSFC / Wyle Information Systems, <joseph.a.hourcle@nasa.gov>



Abstract:

The issues of collaboration, provenance and the ethics of data are not new, but have existed for quite some time. ***The issue is in how to change the attitudes of the scientists*** that these issues are worth their time to deal with, and how the scientists can easily learn what the necessary steps are to ensure that their data can be used by the greater community.

At NASA, there have been changes to the heliophysics data policy which now mentions a need to integrate into the larger data environment [NASA, 2009]. This integration aspect is key, as simple linkages between discrete collections are not enough for effective and efficient reuse of scientific data.

As each scientific mission funded by NASA is novel in some way, there is a belief by many scientists that every data system must be rebuilt from the ground up as well. Although there are often needs to tune systems to meet the needs of the primary investigation, there are limited, if any controls to ensure that the data systems can interoperate with the system of virtual observatories and other cross-discipline efforts being designed and implemented.

We need to find a way to break the 'not-created-here' mentality, and push for PI teams to consider how to support the general science in their discipline when implementing their interfaces. We need to tell them what the requirements are for interfacing with the community search systems, and ***give PI teams a way to get advice on designing and implementing their data system*** in a way that doesn't create roadblocks to the greater community's attempts at using their data.

If we had simple requirements checklists to explain the needs of each discipline, we could give scientists and reviewers an easy way to guage how useful and accesible the system would be. We present a generic checklist developed primarily for file-based feature and event catalogs [Hourclé, 2009], in hopes of inspiring others to develop similar requirements documents for each scientific discipline and to spur discussion of requirements for databases and other larger data systems.

The Catalog Checklist

Although some schools are teaching their science students in the issues of data management, the majority of scientists have no such training. Giving the scientists simple, easy to follow guidance on best practices gives them a tool that they can use to judge how well the results of their work might be received by others. The checklist is not a requirement being forced upon them, but simply ***a list of things that they should consider if they are to publish their results in tabular form***.

This initial version covers the following topics:

- Documentation of the ***catalog***
- Documentation of the ***data*** used for the catalog
- Documentation of the ***records*** in the catalog
- Documentation of the ***attributes*** in the records
- Considerations for ***usability***

We have attempted to ***avoid informatics jargon*** that might turn off the scientist, and hope to ***give short but clear examples of why these items are useful***. As we move to a wiki, it will be possible to give more in-depth explanations.

We hope to keep the checklist short; although there are many items that we could include, we do not want it to grow to an unreasonable size.

Audience Participation:

Have you ever wanted to use someone else's data, and it was just a pain to use?

Could they have done something to improve your use of the data?

Do you have other complaints about other people's data, or ideas of what we should be doing to make re-use of our data easier?

If so, take an index card, ***write down your comments, and pin it up for others to see***. Feel free to ***build on other people's comments***, but please don't remove any, even if you don't agree with them, just put up a counter-argument.

If you're at a loss for ideas, look over the handouts below, and see if they inspire you to comment about something.

Other Checklists

There exist other checklists from the archiving [CRL, 2007], data modeling [Beasley, 2008; Hoberman, 2006], and user interface [W3C, 2008] communities, but they're written for people within the given field. There is a draft from the CCSDS for "Audit and Certification of Trusted Digital Repositories" [2009], but it is lengthy and filled with jargon that can be off-putting to the lay scientist.

The goal of this effort is to educate the scientists, and to frame the issues in ways that would have meaning to them. We hope to ***enable the community to produce better data systems and publish better data***, and thus do better science.

This should be stay a community-based effort, as to be perceived as yet another unfunded mandate will hamper adoption.

For the Future

Ideally, we can try to make sure that the people designing the data systems for future missions are not solely discipline scientists, but are cross-discipline teams including experience in data modeling, archiving and other informatics issues.

As it's unlikely for that to happen with systems already being designed, we can just try to build the best systems that we can set the bar for the rest of the community. So, as a place to record our collective wisdom into what does or does not make a good data system for our field, I've started a wiki at:

http://sciencedata.wikia.com/

I will record the comments posted during this poster session, and hope that others will join and add information that might be of interest to the science informatics community. ***Everyone is welcome to post whatever they might like on the subject***, try to shepard a topic that they're passionate about, or contribute however they can.

References

CCSDS. (2009), "Audit and Certification of Trusted Digital Repositories", CCSDS 652.0-R-1, <http://standards.gsfc.nasa.gov/reviews/ccsds-652.0-r-1/>

CRL. (2007), "Trustworthy repositories audit & certification (TRAC): criteria and checklist", <http://catalog.crl.edu/record=b2212602~S1>

Hourclé, J.A. (2009), "Checklist for Building a Data / Feature / Event Catalog", <http://sdac.virtualsolar.org/catalogs/catalog_checklist>

NASA. (2009), "NASA Heliophysics Science and Data Management Policy: Version 1.1", <http://lwsde.gsfc.nasa.gov/Heliophysics_Data_Policy_2009Apr12.pdf>

Beasley, R. (2008), "Data Modeling Process: Entity Model Definition Checklist", <http://webidesign.com/articles/EntityModelChecklist.pdf>

Hoberman, S. (2006), "The Data Modeling Addict - October 2006", <http://www.tdan.com/view-featured-columns/5473/>

W3C. (2008), "Web Content Accessibility Guidelines (WCAG) Overview", <http://www.w3c.org/WAI/intro/wcag.php>